Section 2

Description of the Sample

This section describes the sample design and selection, the method of estimation, the sampling variability of the estimates, and the methodology of computing confidence intervals.

Domain of Study

The statistics in this report are estimates from a probability sample of unaudited Individual Income Tax Returns, Forms 1040, 1040A, 1040EZ, 1040PC and 1040TEL (including electronic returns) filed by U.S. citizens and residents during Calendar Year 1996.

All returns processed during 1996 were subjected to sampling except tentative and amended returns. Tentative returns were not subjected to sampling because the revised returns may have been sampled later, while amended returns were excluded because the original returns had already been subjected to sampling. A small percentage of returns were not identified as tentative or amended until after sampling. These returns, along with those that contained no income information, were excluded in calculating estimates. This resulted in a small difference between the population total (118,650,252 returns) reported in Table C and the estimated total of all returns (118,218,327) reported in other tables.

The estimates in this report are intended to represent all returns filed for Tax Year 1995. While about 97 percent of the returns processed during Calendar Year 1996 were for Tax Year 1995, a few were for noncalendar years ending during 1995 and 1996, and some were returns for prior years. Returns for prior years were used in place of 1995 returns expected to be received and processed after December 31, 1996. This was done in the belief that the characteristics of returns due, but not yet processed, could best be represented by the returns for previous income years that were processed in 1996.

Sample Design and Selection

The sample design is a stratified probability sample, in which the population of tax returns is classified into subpopulations, called strata, and a sample is randomly selected independently from each stratum. Strata are defined by:

1. Nontaxable with adjusted gross income or expanded income of $200,000 or over and no alternative minimum tax.

2. High combined business and farm total receipts of $50,000,000 or more.

3. Presence or absence of special Forms or Schedules (Form 2555, Form 1116, Form 1040 Schedule C, and Form 1040 Schedule F).

Bonnye Walker and William Wong designed the sample and prepared the text and tables in this section under the direction of Yahia Ahmed, Chief, Mathematical Statistics Section, Statistical Computing Branch.

4. Total gross positive or negative income. Sixty variables are used to derive positive and negative incomes.

5. Potential usefulness of the return for tax policy modeling. Thirty-two variables are used to determine how useful the return is for tax modeling purposes.

Table C shows the population and sample count for each stratum after collapsing some strata with the same sampling rates. (For more details, see references 1 and 2.) The sampling rates range from 0.02 percent to 100 percent.

Tax data processed to the IRS Individual Master File at the Martinsburg Computing Center during Calendar Year 1996 were used to assign each taxpayer's record to the appropriate stratum and to determine whether or not the record should be included in the sample. Records are selected for the sample either if they possess certain combinations of the four ending digits of the social security number, or if their ending five digits of an eleven-digit number generated by a mathematical transformation of the SSN is less than or equal to the stratum sampling rate times 100,000 (see reference 3).

Data Capture and Cleaning

Data capture for the SOI sample begins with the designation of a sample of administrative records. While the sample was being selected, the process was continually monitored for sample selection and data collection errors. In addition, a small subsample of returns was selected and independently reviewed, analyzed, and processed for a quality evaluation.

The administrative data and controlling information for each record designated for this sample was loaded onto an online database at the Cincinnati Service Center. Computer data for the selected administrative records were then used to identify inconsistencies, questionable values, and missing values as well as any additional variables that an editor needed to extract for each record. The editors use a hardcopy of the taxpayer's return to enter the required information onto the online system.

After the completion of service center review, data were further validated, tested, and balanced at the Detroit Computing Center. Adjustments and imputations for selected fields were used to make each record internally consistent, and the data were then tabulated. Finally, prior to publication, all statistics and tables were reviewed for accuracy and reasonableness in light of provisions of the tax law, taxpayer reporting variations and limitations, economic conditions, and comparability with other statistical series.

Some returns designated for the sample were not available for SOI processing because other areas of IRS needed the return at the same time. For Tax Year 1995, 0.23 percent of the sample returns were unavailable.

Method of Estimation

Weights were obtained by dividing the population count of returns in a stratum by the number of sample returns for that stratum. The weights were adjusted to correct for misclassified returns. These weights were applied to the sample data to produce all of the estimates in this report.

Sampling Variability and Confidence Intervals

The sample used in this study is one of a large number of samples that could have been selected using the same sample design. The estimates calculated from these different samples would vary. The standard error (SE) of an estimate is a measure of the variation among the estimates from the possible samples and, thus, is a measure of the precision with which an estimate from a particular sample approximates the average of the estimates calculated from all possible samples.

The standard error may be expressed as a percentage of the value being estimated. This ratio is called the coefficient of variation (CV). Table 1.4 CV contains estimated CV's for the estimates included in Table 1.4 of this report.

The sample estimate and an estimate of its standard error permit the construction of interval estimates with prescribed confidence that the interval includes the population value. If all possible samples were selected under essentially the same conditions and an estimate and its estimated standard error

were calculated from each sample, then:

1. About 68 percent of the intervals from one standard error below the estimate to one standard error above the estimate would include the population value. This is a 68 percent confidence interval.

2. About 95 percent of the intervals from two standard errors below the estimate to two standard errors above the estimate would include the population value. This is a 95 percent confidence interval.

For example, from Table 1.4, the amount estimate for State Income Tax Refunds, X, is $12.24 billion, and its related coefficient of variation, CV(X), is 1.21 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= (\$12.24 \times 10^9) \bullet (0.0121) \\ &= \$0.148 \text{ billion} \end{aligned}$$

The p percent confidence interval is calculated using the formula:

$$X \pm z \bullet SE(X)$$

where z takes the value 1, 2, or 3 when p is 68, 95, or 99, respectively. Based on these data, the 68 percent confidence interval is from $12.091 billion to $12.388 billion, and the 95 percent confidence interval is from $11.943 billion to $12.536 billion.

Table Presentation

Whenever a weighted frequency is less than 3, the estimate and its corresponding amount are combined or deleted in order to avoid disclosure of information for specific taxpayers. (The combined or deleted data, if any, are included in the corresponding column totals.) These combinations and deletions are indicated by a double asterisk (**). Estimates based on less than 10 sampled returns are considered to be unreliable. These estimates are noted by a single asterisk (*) to the left of the data unless all of the sampled returns are selected with certainty (at the 100 percent rate).

In the tables, a dash (- or --) in place of a frequency or an amount indicates that either no returns in the population had the characteristic or the characteristic was so rare that it did not appear on any of the sampled returns.

References

[1] Hostetter, S., Czajka, J. L., Schirm, A. L., and O'Conor, K. (1990), "Choosing the Appropriate Income Classifier for Economic Tax Modeling," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 419-424.

[2] Schirm, A. L., and Czajka, J. L. (1991), "Alternative Designs for a Cross-Sectional Sample of Individual Tax Returns: the Old and the New," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 163-168.

[3] Harte, J.M. (1986), "Some Mathematical and Statistical Aspects of the transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 603-608.

SOURCE:IRS, Statistics of Income, Individual Income Tax Returns 1995, Publication 1304, Revision 3-98.

Table C.—Number of Individual Income Tax Returns in the Population and Sample by Sampling Strata for

Description of the sample strata

Grand total ....

Form 1040 returns only with adjusted gross income or expanded income of $200,000 and over, with no income tax after credits and no additional tax for tax preferences, total

Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total

Other Returns, total ....

Description of the sample strata	Degree of interest [3]	Number of returns by type of form attached							
		Form 1040, with Form 1116 Or Form 2555		Form 1040, with Schedule C but without Form 1116 or Form 2555		Form 1040, with Schedule F but without Form 1116 or Form 2555		All other returns	
		Population Counts	Sample counts	Population counts	Sample counts	Population counts	Sample counts	Population counts	Sam coun
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Total		**1,550,534**	**17,297**	**16,331,658**	**30,601**	**1,639,817**	**4,107**	**99,125,672**	**6**
Negative Income									
$10,000,000 or more	All	102	102	568	568	85	85	851	
$5,000,000 under $10,000,000	All	79	79	688	688	123	123	828	
$2,000,000 under $5,000,000	All	338	99	2,884	851	543	155	3,012	
$1,000,000 under $2,000,000	All	665	89	6,085	826	1,370	191	5,875	
$500,000 under $1,000,000	All	1,499	40	16,135	463	4,112	106	13,291	
$250,000 under $500,000	All	**	**	** 41,864	** 366	10,717	90	28,777	
$120,000 under $250,000	All	**	**	** 88,956	** 352	19,911	82	60,418	
$60,000 under $120,000	All	**	**	** 125,822	** 295	21,707	48	90,916	
Under $60,000	All	**	**	** 353,230	** 298	43,918	41	394,058	
Positive Income									
Under $30,000	1	--	--	--	--	--	--	26,425,079	
Under $30,000	2	81,883	31	1,679,955	549	127,325	48	28,300,200	
Under $30,000	3 - 4	114,708	109	3,194,129	2,481	198,955	173	5,510,279	
$30,000 under $60,000	1 - 2	118,649	43	1,722,822	553	206,357	73	19,922,265	
$30,000 under $60,000	3 - 4	173,838	160	3,170,139	2,782	290,543	255	4,523,998	
$60,000 under $120,000	1 - 3	240,416	81	1,848,472	646	259,105	84	9,490,132	
$60,000 under $120,000	4	194,613	181	2,082,670	2,092	181,382	183	1,760,844	
$120,000 under $250,000	1 - 3	152,367	187	433,121	556	122,903	168	1,318,350	
$120,000 under $250,000	4	180,538	505	990,957	2,784	66,339	160	694,100	
$250,000 under $500,000	All	167,959	1,125	423,066	2,718	62,577	390	415,403	
$500,000 under $1,000,000	All	75,634	1,823	112,869	2,698	15,975	383	116,394	
$1,000,000 under $2,000,000	All	28,988	3,355	26,587	3,122	4,176	505	34,543	
$2,000,000 under $5,000,000	All	13,166	4,196	8,439	2,713	1,352	422	12,546	
$5,000,000 under $10,000,000	All	3,200	3,200	1,548	1,548	246	246	2,388	
$10,000,000 or more	All	1,892	1,892	652	652	96	96	1,125	

[1] This population includes an estimated 431,925 returns that were excluded from other tables in this report because they contained no income information or represented amended or tenta

[2] This population includes 159 Form 1040 returns that were misclassified because of bad data collected during revenue processing.

[3] Each population member is assigned a degree of interest based on how useful it is for tax modeling purposes. Degree of interest ranges from one (1) to four (4), with a one being assigne interesting, and a four being assigned to those that are the most interesting. 'All' refers to income classes for which returns with all four degrees of interest are assigned.

** Data combined.